STATISTICS IN COST PREDICTION PROBLEMS AND POTENTIAL

HARRY PICCARIELLO

Based Upon Presentation to DoD Cost Research Symposium
March 3, 1966

TP 66-5

Office of the Assistant Secretary of Defense (Systems Analysis)

STATISTICS IN COST PREDICTION PROBLEMS AND POTENTIAL

INTRODUCTION

I am going to discuss the application of statistical techniques for determining cost estimates (predictors) in particular:

a) why the use of statistics.

b) problems associated with their use and

c) some possible approaches which avoid the problems or may lead to a solution of these problems. I am referring to statistics here in the technical sense, that is, that branch of applied mathematics which is founded on the theory of probability and commonly called mathematical statistics.

WHY THE USE OF STATISTICS

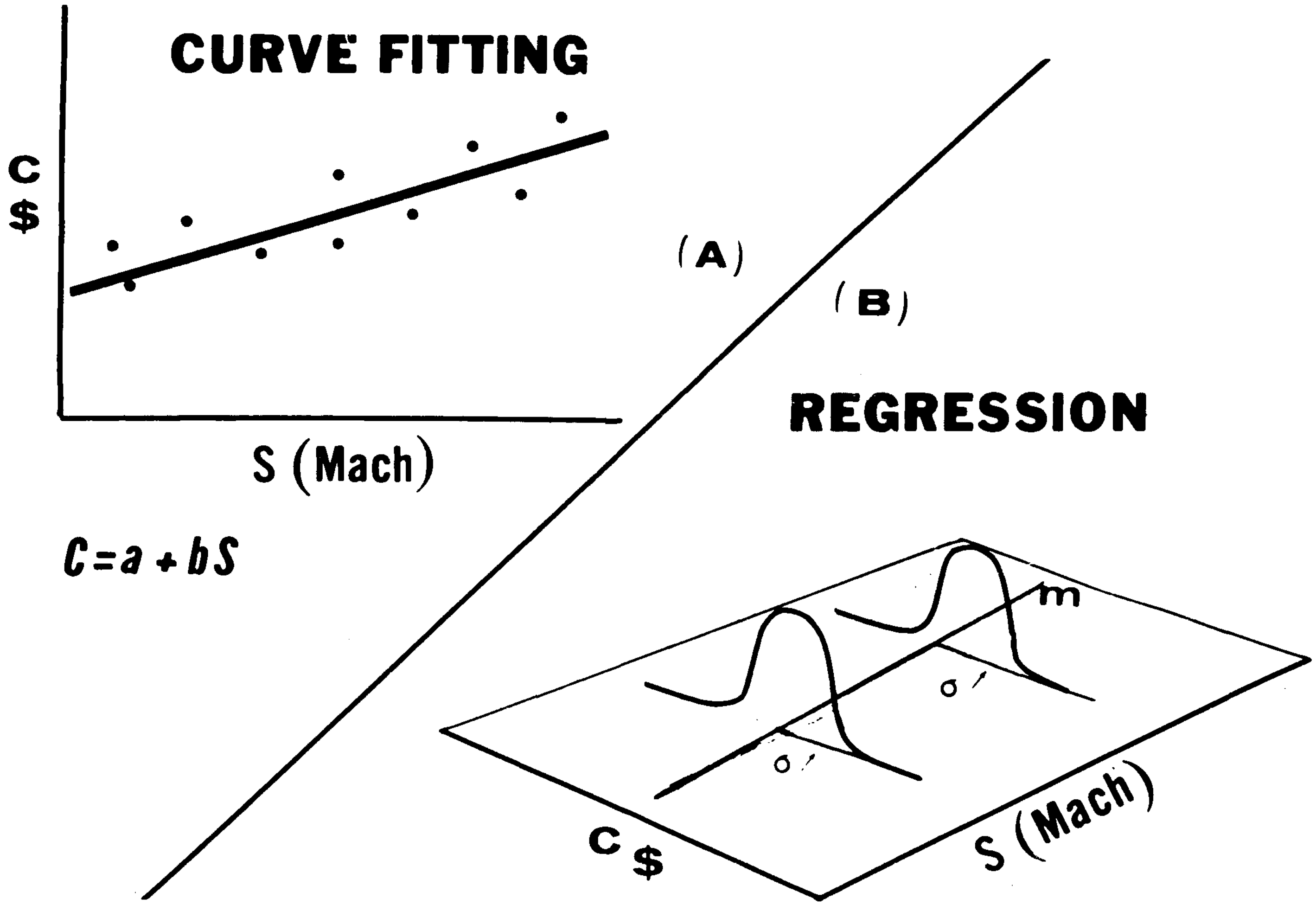
Suppose for the moment that we forget statistics, and we wish to determine a method for estimating the cost of a new aircraft. We have at our disposal cost data on aircraft already built. We assume that the costs are related to certain physical and performance variables. We wish now to determine the unknown relationship between costs and the variables. Assuming that the unknown relationship can be approximated by a function which is linear in these variables (note, however, that we are not restricted to linear functions), the best straight line fit to the data can be obtained. The fit is best in the sense that the sum of the squares of the deviations of the aircraft costs from the desired line is minimized; that this is the least squares line. This process is called curve fitting, and we observe that in carrying it out, there was no reference made to probability distributions, random variables, etc

CHART 1(A) shows the results of applying the foregoing to a sample of 10 aircraft costs at a particular unit (say 100) and where for simplicity we assumed the only variable is speed

Suppose we now introduce the statistical approach, in particular normal regression theory. CHART 1(B) Aircraft costs at the 100th unit are now assumed to be random variables, normally distributed about its mean The mean is a linear combination of performance variables (in this example, speed only) with a constant variance. σ^2 For purposes of estimating, we use the mean as the best estimate of the cost. In deriving the "best" estimates, or minimum variance estimates for the constant a and coefficient of "S" in the equation for the mean ($m = a + bs$) we will find that the derived line of means is identical with the least mean square line of Chart 1.A Why then all the statistical jargon? The reason is that in the former case we can only make statements about how well the line fits the data or approximates the function whereas in the second case we can make probability statements of how well the estimator will predict. We can make statements such as "the cost of the new aircraft will be between x and y with a probability P " That is, considering the problem in statistical terms is an attempt to provide measures of how well the method predicts

With the proliferation of equations purporting to estimate the same things, all coming up with different answers and all claiming excellent fits to the data the need for a way to demonstrate how well the various methods predict becomes apparent Here then is the major problem we are faced with - to find means for demonstrating how well a method or equation predicts Statistics may furnish some answers but there are problems.

Chart 1



PROBLEMS

The first problem is one arising from the use of normal regression theory.

1 Normal Regression Analysis Normal Regression theory is used almost exclusively in deriving CER's. The normality assumption is the least important aspect of the theory in limiting its application to cost estimating. What is important is that the estimate of the cost variable is the mean of the distribution and this mean is linear in functions of the performance variables. The expression for the mean m is, as shown in chart 2, given by:

$$m = a_o + a_1 f_1 \; (S.W.g \; \ldots) + a_2 f_2 \; (S,W,g, \ldots) + \ldots$$

where:

$S, W\ g.\ \ldots$ = various explanatory variables

f_1 = some form of the explanatory variables

a_1 = the coefficients of each form

In applying the theory a selection is made for the f_i, for example

$f_1\ (S.W,g..) = S$

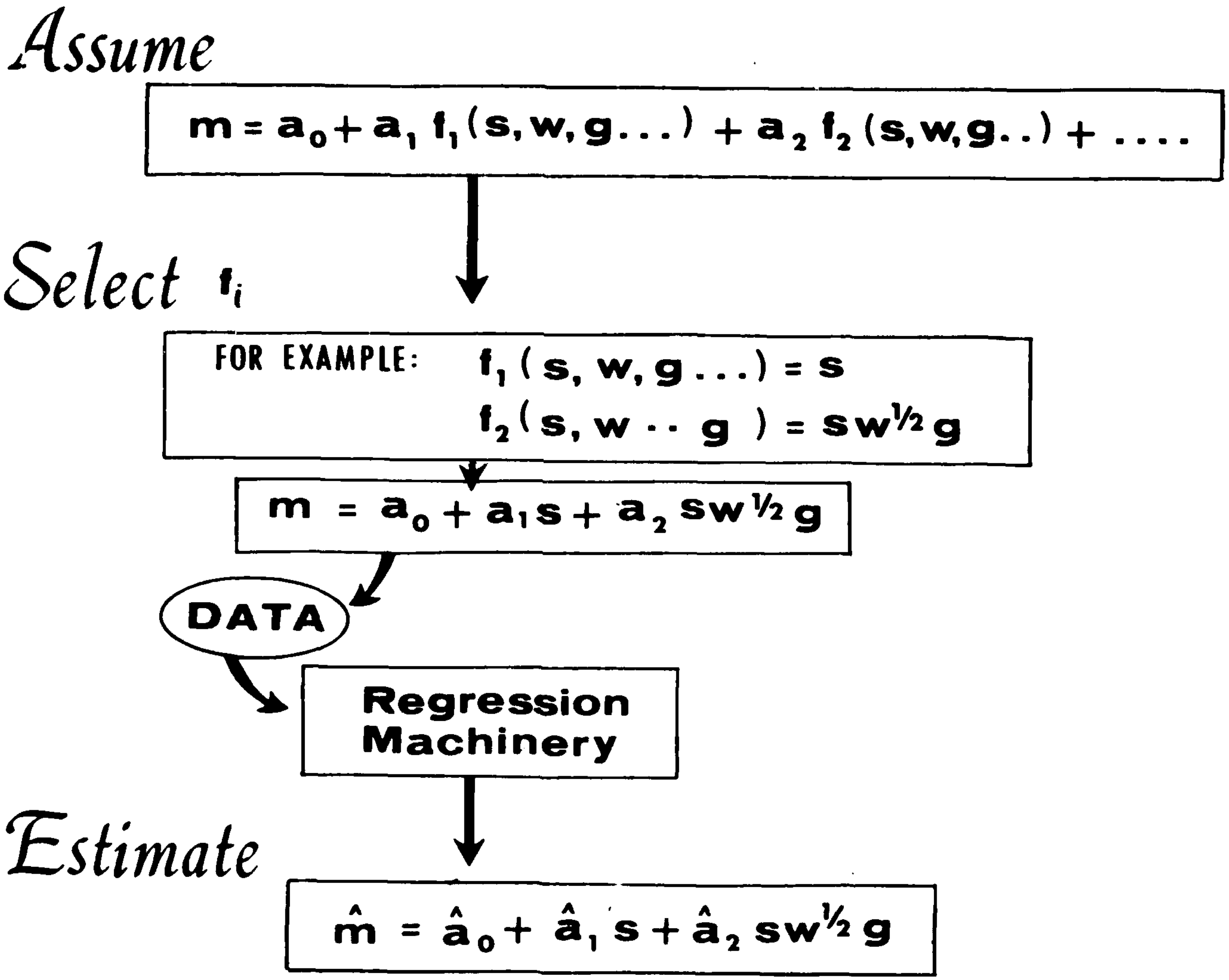
$f_2\ (S\ W,g) = SW^{\frac{1}{2}}$

From the sample data the regression machinery then churns out estimates for the a's which in turn provide an estimate $\hat{m}$ for m:

$$\hat{m} = \hat{a}_o + \hat{a}_1 S + \hat{a}_2 SW^{\frac{1}{2}}$$

Normal theory now allows us to make probability statements about the deviation of a cost to be predicted from its estimate $\hat{m}$ given the f_i are correct However, the theory says nothing about the unknowable and possible major errors in selection of the f_i. It appears that no matter how you select these, if you take enough of them the residual variance can be made quite small and in

Chart 2



Assume

$m = a_0 + a_1 f_1(s, w, g\ldots) + a_2 f_2(s, w, g\ldots) + \ldots$

Select f_i

FOR EXAMPLE: $f_1(s, w, g\ldots) = s$

$f_2(s, w \cdot\cdot\ g) = sw^{1/2}g$

$m = a_0 + a_1 s + a_2 sw^{1/2} g$

DATA

Regression Machinery

Estimate

$\hat{m} = \hat{a}_0 + \hat{a}_1 s + \hat{a}_2 sw^{1/2} g$

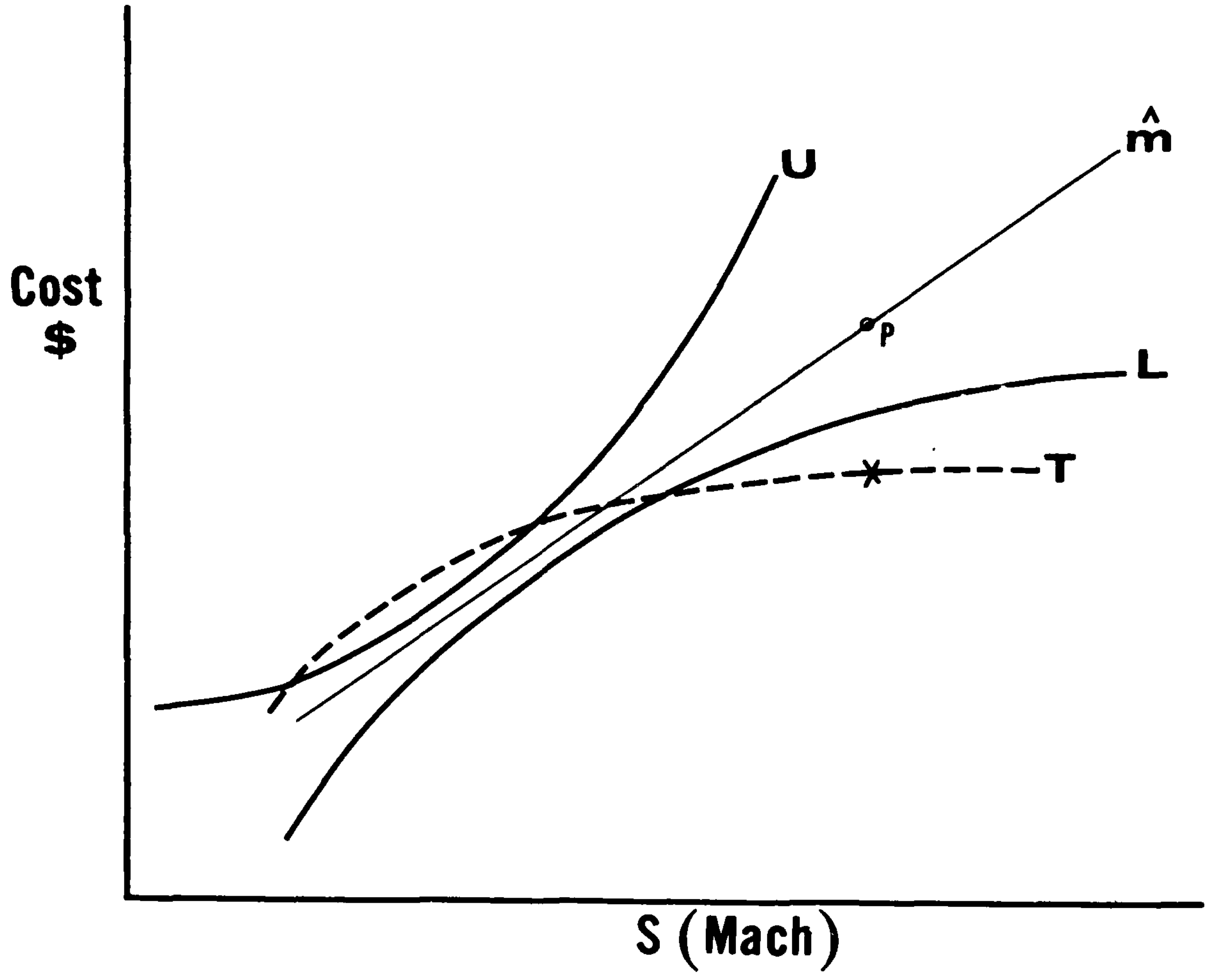
turn the so-called prediction intervals can be made quite small. But the prediction intervals have meaning only if the f_i are known. If they are not known then we can [illegible] no more about our estimate than curve fitting allows us to say and that is "we have a good fit", or "we have a bad fit", and when dealing with small samples this is not very much

On Chart 3 we see [illegible] effect an error in the choice of f_i. can have on our prediction. m is the estimate using the incorrect function. T is the estimate using the true function and U and L are the derived upper and lower 95% prediction intervals about $\hat{m}$. The point P on the line m is the predicted cost of a new aircraft. The theory says that the true cost will lie above or below P and between U and L with a probability of .95. However, the true cost lies outside this interval at the point labelled X. This exemplifies why prediction intervals are questionable if the f_i are unknown.*

2. When total costs are desired, the problem of finding prediction intervals for these totals is compounded by the fact that in general the component costs (e.g. tooling, material, labor, and engineering) are estimated separately. This leads to extremely burdensome problems in determining prediction intervals for the total cost or else requires the introduction of simplifying assumptions which introduce more errors into an already suspect procedure. (For the statistically oriented, this problem is essentially that of determining the convolution of four "t" [illegible]ates all with different weighting factors) When the component cost variables are not independent, the best we can do with certainty is to find upper bounds for the total cost prediction interval. These are almost certain to be so large that they are totally useless. For example the 95% prediction interval for an aircraft might turn out to be 10 million ± 20 million

*These prediction intervals are questioned by some on other grounds, [illegible] small sample available for cost estimating prevents reestimating them by resampling everytime a new prediction is made.

Chart 3



There are other problems but time does not permit a full discussion of them so I will now discuss some other possible approaches which could help in solving these problems

OTHER APPROACHES

1. Historical Simulation

One possible approach to the prediction problem is to use what I call historical simulation.

Evidence to support the worth of a method to predict can be obtained by determining how well it would have predicted in the past The procedure is as follows: Suppose we wish to fit an equation to some data. and suppose for example the equation involved three independent variables. The available data is sorted on time and. say. the three oldest points are used to determine the coefficients The resulting equation is then used to predict all the other points in the data Next the oldest 4 points are used to calculate the coefficients with the resulting equation used to predict all the remaining costs in the data and so on until all points except the last are used to predict the cost of the last We now have "look-see" evidence of how well the method would have performed.

We know that a good predictor should have the property that the coefficients are stable when they are computed using new cost data as these costs become available The technique described above or variations thereof (for example. working backwards using the latest three costs, four costs, etc. to predict the last) could be employed to examine the stability of the coefficients as pertains to the historical data, and if they seem to lack stability. of how the method might be improved

Not much has been done in this area as far as I know and if confronted with small samples not much could be done It may never allow us to make prc.ability statements about the predictor. However it does allow us to look anu see how given methods would have predicted. This in turn could reasonably be used as a criterion for choosing between alternative predictive methods.

2. Direct Estimation of Total Cost.

The following remark pertains to the problem of prediction intervals for total costs. This problem is the simplest to solve. The solution is to estimate the totals directly Summing component cost estimates which are functions of some set of variables cannot improve on a direct estimate of the total which can be taken as a function of the same variables. Although this may not be immediately obvious, this is easily provable.

With a total estimate in hand the problem of computing prediction intervals for the total becomes simple There, no doubt, are good reasons for wanting component costs estimators and nothing precludes our obtaining them for these reasons. However, in my opinion, estimating total costs is not one of these reasons.

3. Selection of Functional Forms.

As I pointed out before, prediction intervals derived from normal regression theory have little meaning if the f_i in the expression for the mean are unknown. I will discuss now an experimental approach that I have tried recently to overcome some of the problems discussed earli

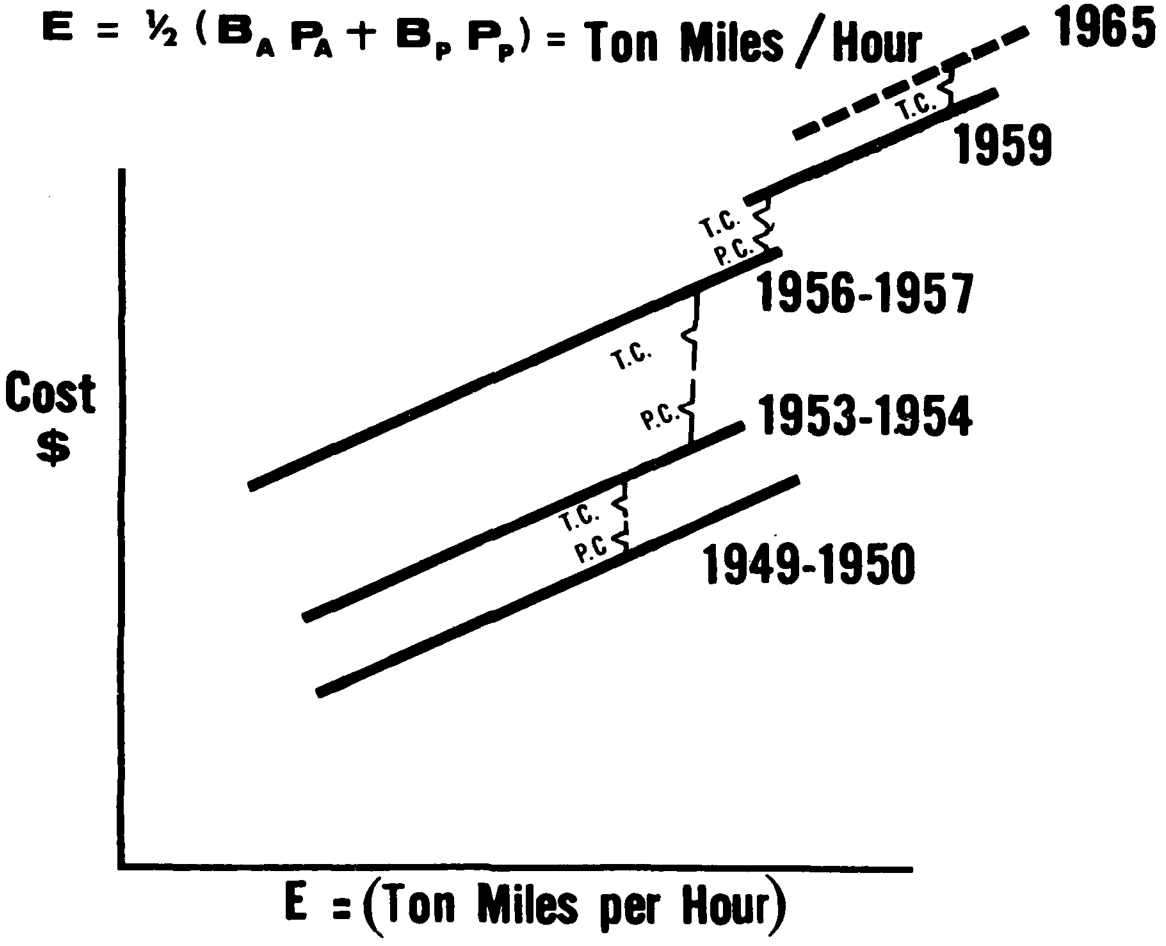
If we can at least assign some heuristic justification to our choice of f_1 then we may have a qualitative criterion for estimating predictive capability between alternative methods which are all good curve fitters. The criterion which I have assumed is that methods which lack a logical justification would be less credible. One possibility might be to relate costs to a combination of certain performance variables which provide a meaningful measure of capability (i.e. what is purchased).

This approach was taken in our office in an experimental effort to develop cost CER for transport cargo aircraft. This project attempted to both utilize a functional form which relates cost to capability and to estimate total flyaway cost directly. The assumed functional relation was $C_N = A(t)E^B$ where E is the arithmetic average of the product of block speed in knots and payload in short tons over Atlantic and Pacific critical legs. A(t) is some function of time t and B is a constant. The sample data was stratified by time periods, and the actual costs were plotted against capability for data in each time period. The resulting curves are shown in Chart 4. The distance between these parallel lines measured along the cost axis is the value of the time dependent cost variation. The large jump that occurs between the 1953-1954 and 1956-1957 time periods. could be accounted for by the fact that all the data points for the first two time periods represented conventional propeller aircraft and all those for the later time periods represented turbo-prop or jet aircraft. In order to use this method for prediction we need to make assumptions about the nature of this time dependent cost variation and how to handle it in making predictions. One possibility is that this total variation is caused by two factors working

Chart 4

$$C_N = A(t)\,E^B$$

$$E = \tfrac{1}{2}\,(B_A P_A + B_p P_p) = \text{Ton Miles / Hour}$$



simultaneously One factor is labor and material price level changes. The other factor is changing labor mixes and material compositions of aircraft with the same capability but built in different time periods and hence in different ways (for example relatively low performance aircraft have been built recently with honeycomb sections whereas many years ago a less expensive method would have been employed). If we assume that this total variation observed in the last period (approximately 3% per year) continues in the same manner until 1965, then predictions for the costs of a new aircraft in 1965 can be read off the 1959 line and then adjusted by a factor of $(1.03)^6$.

Another possiblity is based on the assumption that price level changes can be removed from the total time dependent cost variation by application of a price index. What remains is, for lack of a better term, called technological cost changes. The entities labelled P.C. and T.C. on Chart 4 are the assumed price level and technological change components, respectively of this cost variation. The line labelled 1965 is the result of projecting to 1965 only the technological component of the variation. Predictions for future aircraft can be read off this line and are in terms of 1959 dollars and 1965 technology. Application of an appropriate price index would then be applied to express the costs in 1965 dollars. I am not going to discuss here the worth of this method as a predictor but rather as an example of what might be done and to point out the advantages and disadvantages of this approach.

Advantages:

(1) Costs are related to capability - that is to something which is a meaningful measure of what we are purchasing.

(2) Since costs are assumed to be a function only of time and capability we can, by stratifying by time, eliminate the need to guess the functional form of the time dependent cost variation. It is then possible to examine how costs vary with capability. Each time period furnishes additional evidence as to the correctness of our assumption, and the more time periods we have the more credibility we would place in the method to predict.

(3) By fixing capability it is then possible to make inferences as to the nature of the time dependent cost variation and hence how to handle this in making predictions

Disadvantages:

(1) All the cost generating properties cannot be accounted for by this single measure of capability (for example, reliability)

(2) It is not reasonable to assume that all cost variations not due to capability can be accounted for by time. However, many of the non-capability type qualities an aircraft acquires are a reflection of the state of technology at the time the aircraft is built.

(3) It is not easy and may not be feasible in some cases, to obtain measures of capability

(4) Finally, it is recognized that this approach would limit the sample for two reasons (1) the procedure calls for stratification by mission and (2) the procedure calls for stratification by time periods. However, if it allows us to even get a partial handle on the prediction problem, whatever penalty we pay in reduced sample size is well worth it.

SUMMING UP

The business of cost estimating is prediction, but unfortunately we do not know when we have a good predictor. Statistics as yet does not furnish us an answer to this problem. However such things as historical simulation, relating costs to capability, or other techniques which provide "look see" evidence of the method to predict, could provide some partial answers by, at the very least, establishing reasonable criteria for choosing among alternative methods.